UMVA AI

Cheapest AI Inference & AI-Powered IDE

Pay-As-You-Go. No Subscriptions.

Starting at $0.5/M tokens • 1M Token Context • 384K Output • Native IDE for macOS/Windows/Linux

www.umva.us

San Francisco, CA & Masaka, Kigali, Rwanda

Table of Contents

Table of Contents 2

1. Overview of UMVA AI 5

    Key Highlights 5

2. Models & AI Inference 5

    Model Catalog 6

    Thinking / Reasoning 6

    Why We Are the Cheapest 6

3. UMVA AI IDE 6

    Key Features 6

    System Requirements 7

    Getting Started 7

4. Why Choose UMVA AI 7

5. Pricing & Billing 8

    How Billing Works 8

    Wallet System 9

    Safety Cap 9

    Token Pricing 9

    Real Usage Examples 9

    Fee Comparison: UMVA AI vs Competitors 9

6. Payment Options 10

    Card Payments 10

    Cryptocurrency Payments 10

    Bank Transfer 10

    Payment Method Comparison 10

7. API Documentation 11

    7.1 Authentication 11

    7.2 API Endpoints 11

    7.3 OpenAI-Compatible API 11

    7.4 Anthropic-Compatible API 13

    7.5 Available Models 14

7.6 Request Body Parameters 14
8. IDE Integrations 15
9. Use Cases 15
10. Comparison vs Competitors 17
Feature Comparison 17
Real Cost Comparison by Usage 18
Key Differences at a Glance 18
11. Enterprise & Teams 19
Team Features 19
Enterprise Plan 19
Who UMVA AI Is For 20
12. Security & Privacy 20
Data Handling & Privacy 20
Infrastructure Security 21
API Security Best Practices 21
Compliance & Certifications 21
13. Changelog 21
Version 1.0.0 (Released June 2025) — Latest 21
Version 0.9.0 (Released May 2025) — Beta 22
Upcoming Features 22
14. Frequently Asked Questions 22
General 22
Pricing & Billing 22
Technical 23
IDE & Integrations 23
Security & Privacy 23
Payments 24
15. About UMVA AI 24
Our Team 24
Our Values 25
16. Contact Information 25
Quick Links 25

Additional UMVA Services 26
17. About UMVA Ecosystem 27
The Road Ahead: UMVA's Vision for the Future 27

1. Overview of UMVA AI

UMVA AI is a pay-as-you-go AI platform providing the cheapest AI inference for coding and general use. It includes an AI-powered native IDE for macOS, Windows, and Linux, along with API access for integrating AI into any tool or workflow. All text models feature a massive 1 million token context window and 384K max output tokens. There are no subscriptions — you pay only for the tokens you use, starting at just $0.50 per million input tokens.

UMVA AI was built with a single mission: make AI inference affordable and accessible to every developer on the planet. We saw developers being locked into expensive monthly subscriptions, paying for features they barely used. We believed there was a better way — pay only for what you consume, at prices that make sense.

Our team has deep roots in open-source software, web hosting, and developer tooling. We combined our infrastructure expertise with cutting-edge AI technology to build a platform that delivers enterprise-grade AI inference at a fraction of the cost. Today, UMVA AI serves thousands of developers across 50+ countries.

Key Highlights

- Cheapest AI Inference — Models start at just $0.50 per million input tokens — 3-50x cheaper than competitors.
- No Subscriptions — Pay only for tokens you use. No monthly bills, no annual contracts, no commitments.
- Native AI IDE — Free, native IDE for macOS, Windows, and Linux with AI built in. No Electron bloat.
- 1M Token Context — Process entire codebases in a single request. No more context window errors.
- 384K Max Output — Generate very long responses, entire codebases, or complete documents in one request.
- Pay with Cards or Crypto — Visa, Mastercard, Amex, Verve, USDT, USDC, BTC, ETH — no KYC for crypto.
- No Account Required for IDE — Download and use the IDE for free. Pay only for AI inference.
- Privacy First — Your code stays on your machine. Only selected context is sent for AI requests.
- API Access — Full OpenAI and Anthropic-compatible API for integration with any tool.
- 24/7 Support — Reach us via WhatsApp, live chat, or ticketing system.

---

2. Models & AI Inference

UMVA AI offers specialized text models optimized for coding and general tasks. All models are text-only with a 1M token context window and 384K max output tokens. Every model supports thinking/reasoning (DeepSeek-style extended thinking) and tool use (function calling).

Model Catalog

Attribute	UMVA Code Fast (umva-code-fast)	UMVA Code Architect (umva-code-architect)
Input Price	$0.5000 / 1M tokens	$1.5000 / 1M tokens
Output Price	$1.1200 / 1M tokens	$3.3000 / 1M tokens
Context Window	1,000,000 tokens	1,000,000 tokens
Max Output	384,000 tokens	384,000 tokens
Thinking	Yes (low, high, or max)	Yes (low, high, or max)
Tool Use	Yes	Yes
Best For	Fast and versatile for coding, general tasks, analysis, and creative work	Complex coding, architecture, creative work, and deep analysis

Thinking / Reasoning

At UMVA, all requests are non-thinking by default. You must explicitly pass thinking: {"type": "enabled"} to enable reasoning. This differs from some providers where thinking is enabled by default. Reasoning effort can be set to "high" or "max" ("low" and "medium" map to "high", "xhigh" maps to "max").

Why We Are the Cheapest

- No subscription overhead — you pay only for the compute you use
- Text-only models = no wasted tokens on vision, all compute goes to code
- Direct inference — no middleman markup, no enterprise license fees
- Optimized infrastructure with efficient model serving reduces cost per token
- Safety cap of $0.50 per single request — no runaway costs

---

3. UMVA AI IDE

The UMVA AI IDE is a powerful, native AI-powered code editor that runs on macOS, Windows, and Linux. The IDE is completely free — you only pay for AI inference. It is built for native performance with no Electron bloat, providing instant startup and zero lag.

Key Features

- Lightning Fast — Native performance on every platform. Built for speed with instant startup and zero lag.
- AI Agent Built-In — Your personal AI coding agent. Plan, generate, debug, and refactor code with natural language commands.
- Multi-Agent Mode — Run multiple AI agents in parallel across different files and projects.

- Cross-Platform Native — Native apps for macOS (ARM & Intel), Windows (x64 & ARM), and Linux (x64 & ARM).
- Integrated Terminal — Built-in terminal with AI assistance. Run commands, get explanations, and fix errors.
- No Account Required — Download and use the IDE for free. $0.10 is all you need to start.
- Privacy First — Your code stays on your machine. Only selected context is sent for AI requests.
- Extensible — Support for themes, snippets, language servers, and MCP tools.

System Requirements

- macOS 13+ (Ventura) or later
- Windows 10 22H2+ / Windows 11
- Linux: glibc 2.28+, GTK 3.24+
- 4GB RAM minimum / 500MB disk space
- Internet connection for AI features

Getting Started

1. Download and install the IDE for your OS
2. Launch and sign in with your UMVA account
3. Generate an AI API key from your dashboard
4. Deposit $0.10 or more and start coding
5. Use Cmd/Ctrl+K to invoke the AI agent

---

4. Why Choose UMVA AI

UMVA AI is the smartest choice for AI-powered coding. Here is what sets us apart:

Cheapest Prices Anywhere

Starting at just $0.50 per million tokens, our prices are 3-50x cheaper than competitors like Cursor ($15.00/M), Copilot ($10/month flat), or Zed ($20/month).

No Subscriptions, Ever

Unlike Cursor ($20/mo), GitHub Copilot ($10/mo), or Zed ($20/mo), we have zero recurring fees. You pay only for the tokens you use.

Native AI IDE Included

Free, native AI-powered IDE for macOS, Windows, and Linux. Pre-configured with UMVA AI, featuring built-in AI chat, inline code completion, git integration, and terminal.

1M Token Context Window

All our models support a 1 million token context window — one of the largest available. Process entire codebases in a single prompt.

Pay Only for What You Use

Competitors charge $10-$30/month whether you use their service or not. With UMVA AI, your balance never expires.

No Contracts or Lock-In

Our API is OpenAI and Anthropic-compatible. You can switch providers anytime without changing your code.

Your Data Stays Yours

We never train on your code, never store your prompts, and never share your data.

Crypto Payments Accepted

USDT, USDC, Bitcoin, and Ethereum with no KYC required.

Privacy & Anonymity

No KYC required. No credit card needed to start.

Performance

1M token context window with fast inference. Optimized infrastructure.

Developer-Centric

Built by developers, for developers. Native IDE, API-first design, flexible billing.

Sustainability

Energy-efficient, carbon-aware data centers.

---

5. Pricing & Billing

UMVA AI billing is simple, transparent, and fair. No hidden fees, no complicated tiers, no surprises.

How Billing Works

Add Funds — Add money to your UMVA wallet via card or crypto. Minimum $0.10.

Use AI — Every AI request deducts from your balance based on token usage. Real-time cost tracking.

Track Usage — Monitor your spending in real time from your dashboard.

Recharge — Add more funds anytime. Your balance never expires.

Wallet System

Your UMVA wallet is a prepaid digital wallet. You add funds, and we deduct from your balance as you use AI models. There are no monthly bills, no invoices, and no surprises.

- Balance never expires
- Real-time balance tracking
- Transaction history available
- No minimum balance required

Safety Cap

The Safety Cap is an optional spending limit you can set on your account. It prevents you from accidentally depleting your wallet beyond a certain point. Once you hit the cap, AI requests will be paused until you increase the limit or add more funds.

- Fully customizable
- Per-key spending limits
- Instant notifications
- No surprise charges

Token Pricing

Model	Input (per M tokens)	Output (per M tokens)	1M Cost	10M Cost	Max Output
umva-code-fast	$0.5000	$1.1200	$1.6200	$16.2000	384,000
umva-code-architect	$1.5000	$3.3000	$4.8000	$48.0000	384,000

Prices are in USD. Tokens are counted on both input and output.

Real Usage Examples

- Casual Developer (~5M tokens/month): $8.10/month — less than a cup of coffee
- Regular Developer (~50M tokens/month): $81.00/month — 75% less than competitors
- Power Developer (~500M tokens/month): $810.00/month — still less than a single subscription

Fee Comparison: UMVA AI vs Competitors

Fee Type	UMVA AI	Cursor	Copilot	Zed AI
Subscription Fee	$0.00/mo	$20.00/mo	$10.00/mo	$20.00/mo
Cheapest Input (per M)	$0.50	$15.00	$15.00	$15.00
Cheapest Output (per M)	$1.12	$60.00	$60.00	$60.00
Minimum to Start	$0.10	$20.00/mo	$10.00/mo	$20.00/mo
Balance Expiry	Never	N/A (sub)	N/A (sub)	N/A (sub)
Inactivity Fee	$0.00	N/A (sub)	N/A (sub)	N/A (sub)

6. Payment Options

UMVA AI offers flexible payment options to suit every developer worldwide. Pay with credit/debit cards, cryptocurrency, or bank transfers.

Card Payments

Accepted: Visa (credit & debit), Mastercard (credit & debit), American Express (credit), Verve (debit)

- Minimum deposit: $1.00
- Processing: Instant — wallet credited immediately
- Fees: 2.9% + $0.30
- KYC: Not required
- Security: PCI-compliant gateways, card details never stored

Cryptocurrency Payments

Accepted: USDT (TRC20 & BEP20), USDC (Ethereum & Solana), Bitcoin (BTC), Ethereum (ETH)

- Minimum deposit: $1.00 equivalent
- Processing: 1-15 minutes (depending on network confirmations)
- Fees: Network fee only
- KYC: Zero KYC — full anonymity
- Conversion: Converted to USD at current market rate

Bank Transfer

- Minimum deposit: $10.00
- Processing: 1-3 business days
- Fees: May apply depending on bank
- KYC: Required
- Currencies: USD, EUR, GBP

Payment Method Comparison

Feature	Card	Crypto	Bank Transfer
Processing Time	Instant	1-15 min	1-3 days
Minimum Deposit	$1.00	$1.00	$10.00
KYC Required	No	No	Yes
Fees	2.9% + $0.30	Network fee only	May apply
Anonymity	Partial	Full	None

7. API Documentation

UMVA AI provides both an OpenAI-compatible API and an Anthropic-compatible API for AI completions. The API uses Bearer token authentication and supports SSE streaming.

7.1 Authentication

All API requests require authentication via a Bearer token in the Authorization header. You can generate API keys from your UMVA dashboard.

Header Format (OpenAI):

```
Authorization: Bearer umva_ai_YOUR_API_KEY_HERE
Content-Type: application/json
```

Header Format (Anthropic):

```
x-api-key: umva_ai_YOUR_API_KEY_HERE
anthropic-version: 2023-06-01
```

AI API Key vs Sanctum Token

AI API keys (prefix: umva_ai_) are used for LLM completions and usage stats only. Sanctum tokens provide full API access for account management.

Editor Auth Flow (OAuth-like)

The IDE uses a browser-based token exchange. Opening https://umva.net/account?redirect=umva-editor://callback redirects the browser to authorize the editor with no manual copy-pasting.

7.2 API Endpoints

Method	Endpoint	Description
GET	/v1/models	List models (OpenAI-compatible format)
POST	/v1/chat/completions	Chat completions (OpenAI-compatible, SSE streaming)
POST	/v1/completions	Legacy completions (OpenAI-compatible)
POST	/v1/messages	Messages (Anthropic-compatible, SSE streaming)

Primary Base URL: https://umva.net

Fallback Base URL: https://umva.us

The editor retries the fallback URL on network errors or 5xx status.

7.3 OpenAI-Compatible API

Drop-in replacement for OpenAI's /v1/chat/completions — works with any OpenAI SDK or tool. Compatible with DeepSeek, OpenAI, and any SDK designed for the OpenAI protocol.

Simple cURL — Non-thinking (Default)

```
curl -L -X POST https://umva.net/v1/chat/completions \
  -H "Content-Type: application/json" \
  -H "Accept: application/json" \
  -H "Authorization: Bearer ${UMVA_API_KEY}" \
  -d '{
    "model": "umva-code-fast",
    "messages": [
      {"content": "You are a helpful assistant", "role": "system"},
      {"content": "Hi", "role": "user"}
    ],
    "max_tokens": 4096,
    "stream": false,
    "temperature": 1,
    "top_p": 1
  }'
```

Simple cURL — With Thinking Enabled

```
curl -L -X POST https://umva.net/v1/chat/completions \
  -H "Content-Type: application/json" \
  -H "Accept: application/json" \
  -H "Authorization: Bearer ${UMVA_API_KEY}" \
  -d '{
    "model": "umva-code-fast",
    "messages": [
      {"content": "You are a helpful assistant", "role": "system"},
      {"content": "Solve a complex math problem step by step", "role": "user"}
    ],
    "thinking": {"type": "enabled"},
    "reasoning_effort": "high",
    "max_tokens": 4096,
    "stream": false,
    "temperature": 1,
    "top_p": 1
  }'
```

Python (OpenAI SDK)

```
from openai import OpenAI
client = OpenAI(
    base_url="https://umva.net",
    api_key="umva_ai_YOUR_API_KEY_HERE"
)
# Non-thinking (default)
response = client.chat.completions.create(
    model="umva-code-fast",
    messages=[
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Hello!"}
    ]
)
print(response.choices[0].message.content)

# With thinking enabled
response = client.chat.completions.create(
```

```
    model="umva-code-fast",
    messages=[
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Solve this step by step"}
    ],
    extra_body={
        "thinking": {"type": "enabled"},
        "reasoning_effort": "high"
    }
)
print(response.choices[0].message.content)
```

Node.js (OpenAI SDK)

```
import OpenAI from "openai";
const client = new OpenAI({
    baseURL: "https://umva.net",
    apiKey: "umva_ai_YOUR_API_KEY_HERE"
});
const response = await client.chat.completions.create({
    model: "umva-code-fast",
    messages: [
        { role: "system", content: "You are a helpful assistant." },
        { role: "user", content: "Hello!" }
    ]
});
console.log(response.choices[0].message.content);
```

List Models

```
curl https://umva.net/v1/models \
  -H "Authorization: Bearer ${UMVA_API_KEY}" \
  -H "Content-Type: application/json"
```

7.4 Anthropic-Compatible API

Drop-in replacement for Anthropic's /v1/messages API — works with any Anthropic SDK or tool.

Simple cURL — Non-thinking (Default)

```
curl https://umva.net/v1/messages \
  -H "Content-Type: application/json" \
  -H "x-api-key: ${UMVA_API_KEY}" \
  -H "anthropic-version: 2023-06-01" \
  -d '{
    "model": "umva-code-fast",
    "max_tokens": 4096,
    "messages": [
      {"role": "user", "content": "Hello!"}
    ]
  }'
```

Python (Anthropic SDK)

```
from anthropic import Anthropic
client = Anthropic(
    base_url="https://umva.net",
```

```
    api_key="umva_ai_YOUR_API_KEY_HERE"
)
response = client.messages.create(
    model="umva-code-fast",
    max_tokens=4096,
    messages=[{"role": "user", "content": "Hello!"}]
)
print(response.content[0].text)
```

7.5 Available Models

Model ID	Input Price	Output Price	Context	Best For
umva-code-fast	$0.5000/M	$1.1200/M	1,000,000	Fast and versatile for coding, general tasks, analysis, and creative work
umva-code-architect	$1.5000/M	$3.3000/M	1,000,000	Complex coding, architecture, creative work, and deep analysis

7.6 Request Body Parameters

Parameter	Type	Required	Default	Description
model	string	Yes	-	Model ID to use
messages	object[]	Yes	-	List of messages with role, content
thinking	object|null	No	null	{"type": "enabled"} or {"type": "disabled"}
reasoning_effort	string	No	"high"	Values: "high", "max"
max_tokens	int|null	No	model default	Max tokens to generate
response_format	object|null	No	{"type":"text"}	"text" or "json_object"
stop	string|[]|null	No	null	Up to 16 stop sequences
stream	bool|null	No	false	SSE streaming if true
temperature	number|null	No	1	Sampling temperature 0-2
top_p	number|null	No	1	Nucleus sampling
tools	object[]|null	No	null	Tool definitions (function calling). Max 128.
tool_choice	string|obj|null	No	"none"/"auto"	Tool calling behavior
logprobs	bool|null	No	false	Return log probabilities
user_id	string	No	null	Custom user identifier

Note: frequency_penalty and presence_penalty are deprecated and no longer supported.

At UMVA, all requests are non-thinking by default. You must explicitly pass thinking: {"type": "enabled"} to enable reasoning.

---

8. IDE Integrations

UMVA AI works with your favorite editors and IDEs through our OpenAI and Anthropic-compatible API.

VS Code Integration

UMVA AI works with VS Code through the Continue.dev extension. Install the extension, configure it with your UMVA API key and endpoint, and start using UMVA AI models for code completion, chat, and refactoring within VS Code.

JetBrains IDEs

UMVA AI integrates with all JetBrains IDEs (IntelliJ IDEA, PyCharm, WebStorm, GoLand, etc.) through the Continue.dev plugin or by configuring the built-in AI Assistant feature with UMVA AI as a custom provider.

Neovim / Emacs

Configure UMVA AI as your AI provider in Neovim or Emacs using our API endpoints. Compatible with various AI plugins that support OpenAI or Anthropic protocol.

Zed Editor

UMVA AI works with the Zed editor by configuring it as a custom AI provider. Simply point Zed to our API endpoint with your API key.

Custom Integrations

Our API is fully OpenAI and Anthropic-compatible. Any tool, library, or application that works with OpenAI's or Anthropic's API will work with UMVA AI by simply changing the base URL and API key. This includes LangChain, LlamaIndex, and more.

Base URLs for Integrations

- OpenAI format: https://umva.net (Bearer token auth)
- Anthropic format: https://umva.net (x-api-key auth)
- Fallback: https://umva.us

---

9. Use Cases

UMVA AI is versatile enough to handle virtually any coding task. With the cheapest AI inference on the market, you can experiment freely without worrying about costs.

9.1 Code Generation

Generate boilerplate code, entire functions, classes, and modules from natural language descriptions. UMVA AI excels at understanding context and producing production-ready code in any programming language. With our 1M token context, you can provide extensive codebase context for more accurate

generation. Supported languages include Python, JavaScript, TypeScript, Go, Rust, Java, C++, C#, Ruby, PHP, Swift, Kotlin, and dozens more.

9.2 Debugging & Troubleshooting

Paste error messages, stack traces, or problematic code and get instant analysis. UMVA AI can identify bugs, suggest fixes, explain error messages, and help you debug complex issues. Save hours of troubleshooting with AI-assisted debugging that understands your entire codebase context.

9.3 Code Refactoring

Refactor legacy code, modernize old codebases, split monoliths, and improve code quality. The umva-code-architect model with extended thinking is specifically designed for large-scale refactoring tasks. It can analyze entire codebases and propose architecture improvements, performance optimizations, and design pattern migrations.

9.4 Learning & Education

Learn new programming languages, frameworks, and concepts with AI-powered explanations. Ask questions, get code examples, and understand complex topics with personalized tutoring. With pay-as-you-go pricing, it is the most affordable way to learn coding with AI assistance.

9.5 Documentation

Generate comprehensive documentation for your code — API docs, README files, inline comments, and technical guides. UMVA AI can analyze your codebase and produce clear, well-structured documentation. With 1M token context, it can document entire projects in a single pass.

9.6 Data Analysis & Processing

Process and analyze data with AI assistance. Generate SQL queries, analyze CSV files, create data visualizations, and interpret results. The 1M token context window allows processing of large datasets and documents without splitting.

9.7 Code Review

Automate code reviews with AI. Submit pull requests or code diffs for analysis, get suggestions for improvements, bug detection, and best practice recommendations.

9.8 Test Generation

Generate unit tests, integration tests, and test suites automatically. Analyze function signatures and generate comprehensive test cases covering edge cases, error conditions, and normal operation.

9.9 Code Translation

Translate code between programming languages. Convert Python to JavaScript, Java to Kotlin, Go to Rust, or any language combination. Maintain logic and structure while adapting to target language idioms.

9.10 Shell Scripting

Generate bash, PowerShell, and shell scripts for automation, deployment, system administration, and DevOps tasks. Describe what you need in plain English and get ready-to-run scripts.

9.11 Database Design

Design database schemas, generate SQL migrations, create ER diagrams from descriptions, and optimize queries. Get expert-level database design assistance for any SQL dialect including PostgreSQL, MySQL, SQLite, SQL Server, and more.

9.12 DevOps & Infrastructure

Generate Dockerfiles, Kubernetes manifests, CI/CD pipeline configurations (GitHub Actions, GitLab CI, Jenkins), Terraform scripts, and cloud infrastructure code (AWS, GCP, Azure). Describe your infrastructure needs in plain English.

---

10. Comparison vs Competitors

UMVA AI offers the cheapest AI inference on the market with the most flexible pricing model.

Feature Comparison

Feature	UMVA AI	Cursor	Copilot	Zed AI	Windsurf	Claude Code
Pricing Model	Pay-as-you-go	$20/mo	$10/mo	$20/mo	$15/mo	$20/mo
Minimum to Start	$0.10	$20/mo	$10/mo	$20/mo	$15/mo	$20/mo
Cheapest Input	$0.50	$15.00	$15.00	$15.00	$15.00	$3.00
Cheapest Output	$1.12	$60.00	$60.00	$60.00	$60.00	$15.00
Context Window	1,000,000	Limited	Limited	Limited	Limited	200K
Max Output	384,000	8K-16K	4K-16K	8K-16K	8K-16K	8K-16K
Native IDE	Free ✓	✗	VS Code ext	✗	✗	CLI tool
Train on Code	Never	Opt-out	Opt-out	Opt-out	Opt-out	Opt-out
API Available	Yes ✓	✗	✗	Yes ✓	✗	✗
No Subscription	Yes ✓	✗	✗	✗	✗	✗
Crypto Payments	Yes ✓	✗	✗	✗	✗	✗
No KYC Required	Yes ✓	✗	✗	✗	✗	✗
Thinking/Reasoning	Yes ✓	✗	✗	✗	✗	Yes ✓
Tool Use	Yes ✓	Limited	Limited	✗	✗	Yes ✓

Real Cost Comparison by Usage

Casual User (~10M tokens/month input)

- UMVA AI: $5.00
- Cursor / Zed / Windsurf: $20/mo
- Copilot: $10/mo
- Claude Code: $20/mo

Save $7-$17/mo with UMVA

Regular User (~100M tokens/month input)

- UMVA AI: $50.00
- Cursor / Zed / Windsurf: $20/mo
- Copilot: $10/mo
- Claude Code: $20/mo

UMVA costs more at high usage, but you never pay for idle months

Flexible / Occasional User (skip a month)

- UMVA AI idle month: $0.00
- Cursor / Zed / Windsurf: $20/mo
- Copilot: $10/mo
- Claude Code: $20/mo

Pay zero when you do not use it

Key Differences at a Glance

No Subscription Required:

Every other platform charges $10-20/month whether you use it or not. UMVA AI charges only for actual usage — zero when idle.

Massive 1M Context:

All UMVA models have a 1M token context window. Competitors limit context or charge extra for longer contexts. Process entire codebases in one go.

Cards & Crypto:

Only UMVA accepts crypto payments (USDT, BTC, ETH) alongside credit cards. No other AI coding platform offers this. No KYC required for crypto.

---

11. Enterprise & Teams

UMVA AI for Enterprise provides team management, centralized billing, admin controls, and dedicated support. Give your team access to the cheapest AI inference with the flexibility of pay-as-you-go pricing — no per-seat licenses, no long-term contracts.

Team Features

Team Accounts

Create and manage team accounts with individual API keys for each member. Monitor usage, set spending limits, and control which models each team member can access.

Bulk Billing

One wallet, multiple users. Add funds to a central team wallet and all team members draw from the same balance. No per-seat licenses, no monthly minimums.

Admin Controls

Full administrative dashboard. Manage team members, revoke API keys, set usage caps per user, view detailed usage reports, and configure security policies.

Usage Analytics

Detailed usage analytics per user, per model, and over time. Export reports for internal billing, cost allocation, and capacity planning.

Security Controls

IP whitelisting, API key rotation policies, and role-based access control. Keep your organization's AI usage secure and compliant.

Dedicated Support

Priority support with dedicated account manager. SLA-backed response times, onboarding assistance, and custom integration support.

Enterprise Plan

- Unlimited team members with individual API keys and spending limits
- Centralized billing with a single team wallet
- Admin dashboard with usage analytics and reporting
- IP whitelisting and role-based access control
- Priority email and chat support
- SLA with guaranteed uptime and response times
- Same pay-as-you-go rates: $0.50/M tokens — no per-seat fees, no monthly minimums
- Volume discounts available for high-usage teams

Who UMVA AI Is For

Freelance Developers

No monthly overhead. Pay only for the AI you actually use on client projects.

Startups & SMBs

Keep development costs low with usage-based pricing. No per-seat licenses, no minimum commitments.

Students & Learners

Learn to code with AI assistance for pennies. Pay-as-you-go, no burden of a monthly subscription.

Researchers

Process large datasets and documents with our 1M token context window.

AI Enthusiasts

Build AI-powered applications using our API. Experiment with different models without subscription overhead.

Global Developers

With crypto payments and no KYC, developers from anywhere can access powerful AI without banking restrictions.

---

12. Security & Privacy

At UMVA AI, security and privacy are not afterthoughts — they are foundational. We believe that using AI should never compromise your intellectual property, personal data, or privacy.

Data Handling & Privacy

We Do NOT Train on Your Data:

This is our most important commitment: we never use your prompts, code, or conversations to train or improve our AI models. Your data is processed solely to fulfill your request and is immediately discarded. No training runs. No model fine-tuning. No data harvesting. Period.

No Prompt Storage:

We do not store your prompts or the generated responses on our servers. Once the response is delivered to you, the data is gone. We do not log, archive, or review your conversations.

End-to-End Encryption:

All communication between your IDE or API client and our servers is encrypted using TLS 1.3. Your API keys are hashed and stored securely.

No KYC Required:

You can use UMVA AI without submitting any personal identification. Create an account with just an email, add funds via crypto for complete anonymity.

Infrastructure Security

- Secure Infrastructure: Our servers are hosted in SOC 2 compliant data centers with 24/7 physical security and biometric access controls.
- Threat Detection: We employ real-time monitoring and automated threat detection systems.
- Regular Audits: We conduct regular security audits, penetration testing, and vulnerability assessments.

API Security Best Practices

- API Key Authentication: Keys are hashed using bcrypt and never stored in plaintext. Rotate keys anytime.
- Rate Limiting & Quotas: Prevents abuse and ensures fair usage. Set spending caps on API keys.
- IP Whitelisting: Enterprise accounts can restrict API access to specific IP addresses.
- Usage Logging: Aggregated usage metrics only — no prompt content or generated responses stored.

Compliance & Certifications

Standard	Status
GDPR	Compliant
CCPA	Compliant
SOC 2	Type II
ISO 27001	Certified

---

13. Changelog

Stay up to date with the latest UMVA AI releases, feature updates, improvements, and fixes.

Version 1.0.0 (Released June 2025) — Latest

- Initial Release: Launch of UMVA AI platform with pay-as-you-go AI inference, native IDE, and API access.
- AI Models Launch: Four models: umva-code-fast, umva-code-agent, umva-code-pro, umva-code-architect. All with 1M token context.
- Native IDE v1.0: AI-powered IDE for macOS, Windows, and Linux with built-in git, terminal, AI chat, and inline completion.
- Payment Integration: Card payments (Visa, Mastercard, Amex, Verve) and crypto payments (USDT, USDC, BTC, ETH) with zero KYC.
- OpenAI-Compatible API: Fully OpenAI-compatible API for seamless integration with existing tools and libraries.

- Usage Dashboard: Real-time usage tracking, token consumption analytics, and spending controls with safety cap.

Version 0.9.0 (Released May 2025) — Beta

- Beta Launch: Private beta release with limited users.
- Model Optimization: Performance tuning and latency optimization. Sub-500ms response times.
- Wallet System: Digital wallet with real-time balance tracking and transaction history.
- IDE Prototype: First working prototype of the native AI IDE with basic chat and code completion.

Upcoming Features

- Voice Coding
- Image Generation
- Agent Mode
- Custom Models
- Mobile IDE
- Edge Inference

---

14. Frequently Asked Questions

General

What is UMVA AI?

UMVA AI is a pay-as-you-go AI platform that provides the cheapest AI inference for coding and general use. It includes an AI-powered IDE that runs on macOS, Windows, and Linux, along with API access for integrating AI into your own tools. All text models feature a 1M context window, and you only pay for what you use — no subscriptions, no monthly commitments.

Is UMVA AI free to use?

UMVA AI is pay-as-you-go from the start — no free trial needed because you only pay for what you use. Just add as little as $0.10 to your wallet and start coding immediately. No credit card is required to create an account.

Do I need to create an account?

Yes, you need a free UMVA account to use our AI services. Registration requires only an email address and password. No KYC or identity verification is required.

Pricing & Billing

How does pay-as-you-go pricing work?

You add funds to your UMVA wallet and pay per token as you use AI models. Pricing is based on input and output tokens per million tokens — starting at $0.50. Your usage is deducted from your balance in real time. No monthly fees, no subscription tiers, no expiration on credits.

What are the current model prices?

umva-code-fast: $0.50/M input, $1.12/M output. umva-code-architect: $1.50/M input, $3.30/M output. All models support up to 384,000 output tokens. All prices in USD.

Do my credits expire?

No. Funds in your UMVA wallet never expire. You can keep your balance for as long as you like. No inactivity fees or expiration dates.

Technical

What is the context window size?

All UMVA AI text models feature a 1 million token (1M) context window — one of the largest available from any AI provider.

What is the max output token limit?

All UMVA AI models support up to 384,000 (384K) output tokens per request, significantly higher than most providers (4K-16K).

Is the API OpenAI-compatible?

Yes. UMVA AI provides a fully OpenAI-compatible API. Any client that works with OpenAI's API will work with UMVA AI by simply changing the base URL and API key. It is also Anthropic-compatible.

IDE & Integrations

What is the UMVA AI IDE?

A free, native AI-powered code editor for macOS, Windows, and Linux with built-in AI chat, inline code completion, git integration, and terminal.

Can I use UMVA AI with VS Code?

Yes, through the Continue.dev extension. Configure it with your UMVA API key and endpoint.

Does UMVA AI work with JetBrains IDEs?

Yes, through the Continue.dev plugin or by configuring the built-in AI Assistant with UMVA AI as a custom provider.

Security & Privacy

Do you train on my data?

No. We never use your prompts, code, or conversations to train or improve our AI models. Your data is processed solely to fulfill your request and is immediately discarded.

Do you store my prompts?

No. We do not store your prompts or generated responses. Once the response is delivered, the data is gone.

How are my API keys protected?

API keys are hashed using bcrypt and never stored in plaintext. All communication is encrypted using TLS 1.3.

Payments

What payment methods do you accept?

Visa, Mastercard, American Express, Verve cards, plus cryptocurrency: USDT (TRC20 & BEP20), USDC, Bitcoin, and Ethereum. No KYC required for crypto.

What is the minimum deposit?

$1.00 for card and crypto payments. $10.00 for bank transfers. No maximum limit.

Can I get a refund?

Wallet top-ups are non-refundable, but unused funds remain in your wallet indefinitely with no expiration.

---

15. About UMVA AI

UMVA AI began as a response to the rising costs of AI coding assistants. We saw developers being locked into expensive monthly subscriptions, paying for features they barely used. We believed there was a better way — pay only for what you consume, at prices that make sense.

Our mission is radical affordability, open and accessible AI, privacy-first design, uncompromising performance, and developer-centric tools. Built by developers, for developers, UMVA AI serves thousands of developers across 50+ countries.

Our Team

UMVA Team (Engineering & Operations):

A distributed team of engineers, AI researchers, and open-source contributors spread across the globe, united by a shared mission to democratize AI access.

Open Source Community (Contributors & Partners):

Our IDE and many of our tools are open-source, with contributions from developers around the world.

Support Team (Customer Success):

We pride ourselves on responsive, human support. When you reach out, a real person answers.

Our Values

Radical Affordability

AI should not be a luxury. Starting at $0.50/M tokens — the lowest in the industry.

Open & Accessible

No KYC. No credit card needed to start. Open, well-documented API.

Privacy First

We never train on your data, never store your prompts, never share your information.

Performance

1M token context window with fast inference. Optimized infrastructure.

Developer-Centric

Native IDE, API-first design, flexible billing.

Sustainability

Energy-efficient, carbon-aware data centers.

---

16. Contact Information

Have questions, need support, or want to discuss custom requirements? The UMVA team is available through multiple channels.

Channel	Details
Website	umva.us
UMVA AI Home	umva.us/umva-ai
API Documentation	umva.us/umva-ai/api-docs
Email (Support)	info@umva.us
Phone / Telegram / WhatsApp	+1 (415) 377-2023
Office (USA)	San Francisco, CA, United States
Office (Rwanda)	Masaka, Kigali, Rwanda

Quick Links

- UMVA AI Home: umva.us/umva-ai
- AI Model Pricing: umva.us/umva-ai#tab-pricing
- Download IDE: umva.us/umva-ai#tab-download
- About UMVA AI: umva.us/umva-ai/about

- Why UMVA AI: umva.us/umva-ai/why-us
- vs Competitors: umva.us/umva-ai/comparison
- Changelog: umva.us/umva-ai/changelog
- Enterprise: umva.us/umva-ai/enterprise
- Developer Portal: umva.us/developer
- How Billing Works: umva.us/umva-ai/billing
- Payment Options: umva.us/umva-ai/payment-options
- API Documentation: umva.us/umva-ai/api-docs
- Use Cases: umva.us/umva-ai/use-cases
- IDE Integrations: umva.us/umva-ai/integrations
- FAQ: umva.us/umva-ai/faq
- Security & Privacy: umva.us/umva-ai/security
- Contact Support: umva.us/contact

Additional UMVA Services

UMVA is a comprehensive digital business platform offering: Software Licensing, Scripts Marketplace, SEO Services, SMS & WhatsApp Business API, Email Servers, Domain Registration, Web Hosting (cPanel, VPS, VDS, Reseller), Virtual Cards, and Payment Gateway solutions — all from one account.

Licensing

Authorized reseller for cPanel, WHMCS, JetBackup, LiteSpeed, and other popular server and hosting software. Licenses delivered instantly, with volume discounts available.

Scripts Marketplace

Curated collection of ready-to-deploy scripts — CMS platforms, e-commerce solutions, automation tools, and custom integrations.

SMM & Digital Marketing Lab

Social media marketing services, SEO services (technical optimization, link building, keyword research), and performance tracking.

Communication APIs

SMS and WhatsApp Business APIs for transactional messages, marketing campaigns, and notification alerts. Global reach across 200+ countries.

Email Servers

Professional email hosting with SMTP, IMAP, webmail. SPF, DKIM, DMARC authentication. 99.9% uptime guarantee.

Domain Registration

All major TLDs (.com, .net, .org, .io, .co, .app, country-code TLDs). DNS controls, WHOIS privacy, auto-renewal.

Web Hosting

Shared hosting (cPanel), VPS (managed/unmanaged), VDS, and Reseller hosting. SSD storage, DDoS protection, automated backups.

Virtual Cards

Instant virtual Visa, Mastercard, and Verve cards with no KYC, 5-year validity, 0% FX fees.

Payment Gateway

Merchant payment processing for cards, crypto, and mobile money. REST APIs and plugins for popular e-commerce platforms.

---

17. About UMVA Ecosystem

UMVA is not just an AI platform — it is a comprehensive digital business ecosystem designed to give individuals and businesses everything they need to succeed online. From infrastructure like web hosting and domain registration to financial tools like virtual cards and payment gateways, UMVA provides a unified platform where all your digital services are managed from a single account.

The UMVA ecosystem is built on the principle of accessibility. Just as UMVA AI removes barriers to AI-powered coding, every UMVA service is designed to be accessible without the traditional gatekeepers. No lengthy approval processes, no credit checks, no unnecessary paperwork.

The UMVA vision is a unified digital ecosystem where one account gives you access to everything you need to build, run, and grow an online business. From the infrastructure that hosts your website to the AI that powers your code, every service is designed to work together seamlessly. As UMVA continues to grow, new services and features are added regularly, all accessible from the same familiar dashboard.

The Road Ahead: UMVA's Vision for the Future

UMVA is committed to continuous improvement and expansion of its ecosystem. Upcoming developments for UMVA AI include voice coding, image generation, autonomous agent mode, custom model hosting for enterprises, a mobile IDE, and edge inference nodes for lower latency.

The ultimate goal is to create a truly borderless digital economy where anyone, anywhere, can participate fully in AI-powered development without being limited by their geography, banking status, or access to traditional financial infrastructure. UMVA AI is at the forefront of that journey, and the entire UMVA ecosystem is designed to support it.

© UMVA • All rights reserved

SSL Secured • Encrypted